

# Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data

**Axelsen, Martin Christian; Bak, Nikolaj; Hansen, Lars Kai**

*Published in:*
Proceedings of the 5th International Workshop on Pattern Recognition in NeuroImaging (PRNI 2015)

*Link to article, DOI:*
10.1109/PRNI.2015.20

*Publication date:*
2015

*Document Version*
Peer reviewed version

# Testing Multimodal Integration Hypotheses with Application to Schizophrenia Data

Martin C. Axelsen*†, Nikolaj Bak†‡ and Lars K. Hansen*
*Cog-Sys - DTU Compute, Technical University of Denmark
Kgs. Lyngby, Denmark, Email: maxe@dtu.dk
†Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS),
Psychiatric Centre Glostrup, Mental Health Services, Capital Region, Denmark
‡Center for Neuropsychiatric Schizophrenia Research (CNSR),
Mental Health Services, Capital Region, Denmark

***Abstract*—Multimodal data sets are getting more and more common. Integrating these data sets, the information from each modality can be combined to improve performance in classification problems. Fusion/integration of modalities can be done at several levels. The most appropriate fusion level is related to the conditional dependency between modalities. A varying degree of inter-modality dependency can be present across the modalities. A method for assessing the conditional dependency structure of the modalities and their relationship to intra-modality dependencies in each modality is therefore needed. The aim of the present paper is to propose a method for assessing these inter-modality dependencies. The approach is based on two permutations of an analyzed data set, each exploring different dependencies between and within modalities. The method was tested on the Kaggle MLSP 2014 Schizophrenia Classification Challenge data set which is composed of features from functional magnetic resonance imaging (MRI) and structural MRI. The results support the use of a permutation strategy for testing conditional dependencies between modalities in a multimodal classification problem.**

## I. Introduction

Schizophrenia is a complex disorder with a very heterogenous symptomatology [1]. To assist diagnosis many quantitative techniques including neuroimaging have been proposed although no modality has solved the diagnosis problem yet. Hence, new proposed diagnostic tools typically face a complex multimodal decision challenge. Combining data from several modalities in a classification pipeline is not trivial as this can be done at several levels. Multimodal decision problems are in fact relevant to several fields, and three different levels of integration are applied ranging from the early integration of modalities in data level fusion towards an intermediate integration at the feature or representation level, and finally a late integration often named decision level [2][3]. Hybrid integration is also discussed where data is fused at different levels [2][4].

In a review by Sui et al [5], several multivariate methods of early to intermediate fusion of brain imaging data are discussed. The earlier fusion levels (denoted data fusion) are chosen for the review[5] as these allow for access to potential joint information between the several modalities where later fusion (denoted data integration) preclude the decision model to explore such information.

Later fusion may however lead to simpler models with less parameters to be inferred, hence, potentially less data overfitting and reduced computational complexity. An additional benefit from a late fusion scheme, is that it will be possible to obtain results in cases where one or more modalities are missing. This is of particular interest in complex diagnostic problems where patient conditions can preclude acquisition of data [6].

Conditional independencies can be explored for the identification of the right level of integration for a given data set. Let $\mathbf{y}$ be the relevant decision label, $\mathbf{h}$ latent variables, and $\mathbf{u} = (\mathbf{v}_1, \mathbf{v}_1, ..., \mathbf{v}_J)$ be observed multimodal data for $J$ modalities. Decision theory tells us that the error rate (or more generally the expected loss) is minimized when decisions are based on the posterior probability $p(\mathbf{y}|\mathbf{u})$. To guide our inference procedures, we use Bayes theorem to rewrite

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{u}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{u})}. \tag{1}$$

If the only information shared between the modalities is the decision label, i.e., the modalities are conditionally independent, we obtain

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{y})\prod_{j=1}^{J} p(\mathbf{v}_j|\mathbf{y})}{p(\mathbf{u})}. \tag{2}$$

This allows for a multiplicative combination scheme for the 'Bayesian surprise' as defined in [7], hence, by normalization, of the posterior probability of interest

$$\frac{p(\mathbf{y}|\mathbf{u})}{p(\mathbf{y})} = \prod_{j=1}^{J}\left[\frac{p(\mathbf{y}|\mathbf{v}_j)}{p(\mathbf{y})}\right]\left[\frac{\prod_{j=1}^{J} p(\mathbf{v}_j)}{p(\mathbf{u})}\right]. \tag{3}$$

This is a late fusion scheme as we essentially combine $J$ independent inference pipelines to reach the optimal decision function $p(\mathbf{y}|\mathbf{u})$.

Now, such conditional independence with respect to the diagnosis, $\mathbf{y}$, may be a too strong assumption. Other relevant features known or unknown could create dependencies between the modalities, such as age, gender, or endo-phenotypes. If we denote these variables by $\mathbf{h}$, then by a similar argument, modality independence conditioned on labels and

latent variables, i.e., $p(\mathbf{u}|\mathbf{y},\mathbf{h}) = \prod_{j=1}^{J} p(\mathbf{v}_j|\mathbf{y},\mathbf{h})$, leads to an 'intermediate' level fusion scheme, where a first set of independent pipeline modules identify $\mathbf{h}$.

As the salient dependency structures are unknown, we may simply as a first step (A) explore the relative merits of early, intermediate, and late fusion architectures. As a step towards more detailed understanding, however, we here suggest two additional permutation steps to test dependencies in a given data set. We propose a permutation step designed to increase inter-modality dependency and a step to remove them. The permutation schemes are illustrated in Fig 1. In proposed step (B) we permute the measurement variables to create a randomized set of "pseudo-modalities" each having the same number of variables as in the original measurements. By this operation, possible within modality dependencies are transformed to become dependencies between modalities. If we adapt an early fusion model on such permuted data we clearly expect no difference compared with the early fusion model for case (A). However, when training intermediate or late fusion models a performance drop will inform us that the biases introduced in these simpler architectures are too strong for the task. Our final permutation strategy (C) tests the assumption of conditional independence on labels $\mathbf{y}$. We create a new data set in which we permute the sample indices on the individual modalities among data within a specific $\mathbf{y}$ group, i.e., if we consider a binary decision problem, we randomly mix modality subsamples within the two label groups. Thereby we create a new sample in which any other dependency than that induced by $\mathbf{y}$ has been removed. Under the late fusion hypothesis this step should not decrease performance relative to late fusion in case (A).

## II. Materials and methods

### A. Data

Data was obtained from the Kaggle website (https://www.kaggle.com): "The MLSP 2014 Schizophrenia Classification Challenge" (partially describe in [8]). It consists of 378 features from a functional magnetic resonance imaging (fMRI) paradigm and 32 features from a structural MRI (sMRI) scan from 86 observations (40 schizophrenia patients and 46 healthy controls). The Kaggle challenge was to classify patients vs. controls (binary classification). Only the labeled part of the dataset was used.

### B. Pipeline

In order to investigate the dependencies between modalities, three levels of fusion (early, intermediate, late) were tested. First, input feature selection was performed with the filter method [9]. The ten lowest ranking input features (judged by p-value) from each modality were included. The number of included features were a compromise between the dimensions of the original modalities and our aim to treat the modalities at approximately same footing.

The main non-linear processing step, designed to infer relevant latent features $\mathbf{h}$, consisted of a restricted Boltzmann machine (RBM). The decision step was performed with logistic regression on the binary nodes and the group, see Fig. 2. The restricted Boltzmann machines were modified from the implementation in [10] to accommodate Gaussian distributed visible units. Contrastive divergence as introduced by Hinton [11] is used for learning.
Assuming that the variance of the input data is 1, the updates for the visual and hidden units are then given by

$$\mathbf{h}_{data} = \sigma(\mathbf{b} + \mathbf{v}_{data}\mathbf{w}^T) > 0.5 \tag{4}$$

$$\mathbf{v}_{recon} = \mathbf{a} + \mathbf{h}_{data}\mathbf{w} + \epsilon \tag{5}$$

$$\mathbf{h}_{recon} = \sigma(\mathbf{b} + \mathbf{v}_{recon}\mathbf{w}^T) > 0.5 \tag{6}$$

where $\epsilon$ is unit variance Gaussian white noise, $\sigma$ is a sigmoid function, $\mathbf{a}$ is the bias for the visual units, and $\mathbf{b}$ is the hidden unit bias. Thus, the updates for each variable are

$$\widehat{\mathbf{w}} = \beta\widehat{\mathbf{w}}_{-1} + \alpha((\mathbf{vh})_{data} - (\mathbf{vh})_{recon}) \tag{7}$$

$$\widehat{\mathbf{a}} = \beta\widehat{\mathbf{a}}_{-1} + \alpha(\mathbf{v}_{data} - \mathbf{v}_{recon}) \tag{8}$$

$$\widehat{\mathbf{b}} = \beta\widehat{\mathbf{b}}_{-1} + \alpha(\mathbf{h}_{data} - \mathbf{h}_{recon}) \tag{9}$$

where $\alpha$ is the learning rate, $\beta$ is the momentum and the subscript $(-1)$ denotes the update of the variable from the previous iteration.
Weights were initialised randomly, the bias for the visual units was initialised as $\log\left(\frac{40/86}{1-40/86}\right)$, and the bias for the hidden units was initialised as zeros, all as recommended in [12]. The values for the learning rate ($\alpha = 10^{-2}$), momentum ($\beta = 2^{-1}$) and batch size (10 samples) were found according to this guide as well. Logistic regression was finally used for computing posterior probability outputs with the RBM nodes as input and the group as label.

Grid searches were done to assess the optimal number of hidden nodes in each integration procedure. This was done to ensure that possible differences in performance could not be attributed to model complexity or number of parameters alone.

The aim of the present study was to explore more formally the effects of possible conditional independence between modalities, given the group. Therefore, based on the original data set (data set A), two additional data sets were created as described above by permutations of input features (B) and among observations (C), respectively (see Fig. 1). In the first permutation strategy (data set B), the features from the two modalities, fMRI and sMRI, were mixed, so two new pseudo modalities were created based on the original data. The grid search for this combination was restricted to equal number of hidden nodes by symmetry. In the second permutation strategy (data set C) the sample indices within a given label group were randomly permuted for each modality.

### C. Crossvalidation

The performances of the three levels of integration in each data set were estimated with an 8 fold cross validation procedure on the entire pipeline after preprocessing (feature selection). Learning curves were computed for the best performing number of nodes for early, intermediate and late fusion in data set A in a "leave two out" cross validation. In each fold, a complete learning curve was estimated varying

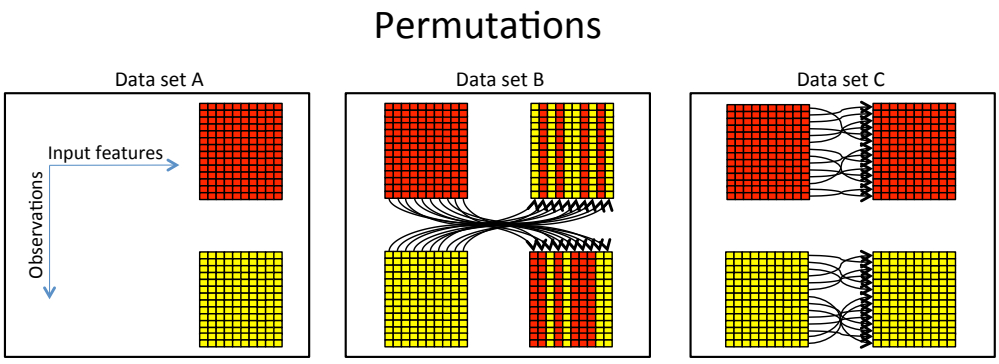


Fig. 1. The three data sets analysed. A is the original set with input features selected using a simple univariate test. In data set B the features are permuted between modalities. In data set C the observations are permuted group wise for one modality.

the size of training data from 2 to 78, but keeping group proportions in training data fixed at 50%. Both the 8 fold- and the "leave two out" cross validation experiments were repeated 300 times each.

## III. Results and Discussion

Generalization performance as assessed by crossvalidation for the different combinations of early, intermediate, and late integration experiments on data set A, B, and C, are shown in table I. The overall best performance is obtained for late integration when based on data sets A or C. For data set B, with presumably high inter-modal dependency, early integration shows the best performance, and is equivalent to the performance seen in early integration model on data set A ($p = 0.7$). For completeness we also tested performance of the models trained on the individual modalities separately, finding somewhat higher test errors.

The results from the grid search experiment on the individual modalities and on the three levels of fusion of data set A are shown in Fig. 3. The red square marks the overall lowest error, which is seen in late fusion with one node for the functional modality and two for the structural. For the early fusion, a peak in error is seen for 2 nodes. An increase at two nodes is also seen for the individual modalities, which could be an indication that in this case, the RBM finds some alternative hidden variables e.g. age or gender. For the late and especially for the intermediate fusion, an increase in error is seen as more nodes are included. The lowest error is in both cases found centred around (2,2) nodes.

Learning curves for early, intermediate, and late data fusion of data set A are seen in Fig. 4. From this it is seen that the hidden unit selection scheme seems to prevent overfitting.

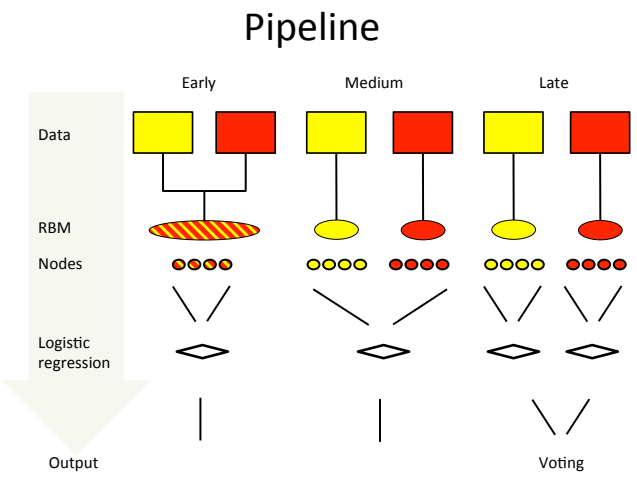


Fig. 2. Early, intermediate, and late fusion pipelines.

TABLE I
Node combination, mean error, and standard error of the mean from individual modalities and for early (E), intermediate (I), and late (L) fusion of data sets A, B, and C. Lowest error is bold for each data set. For E, I, and L, the p-value from a two-sample t-test between the fusion levels are shown. As the comparison is commutative, only the lower triangle of the 3x3 matrix is shown.

| Individual modalities | No Nodes F | No Nodes S | Err | SE | | |
|---|---|---|---|---|---|---|
| Functional | 1 | | 2.101e-01 | 2.517e-04 | | |
| Structural | 1 | | 2.752e-01 | 5.755e-04 | | |
| **Data set A** | **No Nodes F** | **No Nodes S** | **Err** | **SE** | **Stat Diff ($p$) E** | **Stat Diff ($p$) I** |
| Early | 3 | | 1.785e-01 | 1.346e-03 | - | - |
| Intermediate | 1 | 2 | 1.907e-01 | 1.312e-03 | 2e-10 | - |
| Late | 1 | 2 | **1.748e-01** | 1.122e-03 | 4e-02 | 6e-19 |
| **Data set B** | **No Nodes F/S** | **No Nodes F/S** | **Err** | **SE** | **Stat Diff ($p$) E** | **Stat Diff ($p$) I** |
| Early | 3 | | **1.793e-01** | 1.444e-03 | - | - |
| Intermediate | 3 | 3 | 2.092e-01 | 1.805e-03 | 8e-34 | - |
| Late | 4 | 4 | 2.001e-01 | 1.615e-03 | 2e-20 | 2e-04 |
| **Data set C** | **No Nodes F** | **No Nodes S** | **Err** | **SE** | **Stat Diff ($p$) E** | **Stat Diff ($p$) I** |
| Early | 1 | | 1.834e-01 | 1.117e-03 | - | - |
| Intermediate | 1 | 2 | 1.843e-01 | 1.294e-03 | 6e-01 | - |
| Late | 1 | 2 | **1.752e-01** | 1.131e-03 | 4e-07 | 1e-07 |

Rather low dimensional models are chosen, except in the data set B. The grid search results in Fig 3 illustrates that the model is inclined to choosing a simple model.

The results on the original data and the two permuted data sets together present evidence for the basic conditional independence hypothesis: The label patient/control is the strongest link between the two modalities. The primary evidence is the statistically significant (at 5% significance level) improved performance of the late fusion model in data set A. In addition, the fact that the late fusion model on the permuted set C achieves the lowest performance of the three fusion levels (and the same performance as seen in A) supports this hypothesis. The permutation scheme breaks any other dependency structure between the two modalities, though it still maintains performance, and with a model of same complexity (number of hidden RBM units).

The fact that data set B, which presumably has strong inter-modality dependencies, shows poor performance when modelled with intermediate and late fusion architectures, again is evidence that with the given signal-to-noise ratio and sample sizes we are in fact able to detect dependencies, when they exist.

## IV. Conclusion

Our results provide evidence that the proposed permutation strategies can elicit the conditional dependency structure among modalities in a multimodal decision problem. The present work is to the best of our knowledge the first to use such permutation schemes. Future work should be conducted to expand and test the procedures on a broader selection of data sets, and also to further investigate the effects of the feature and model selections. The implications for Schizophrenia diagnosis support should likewise be explored to a greater extend.

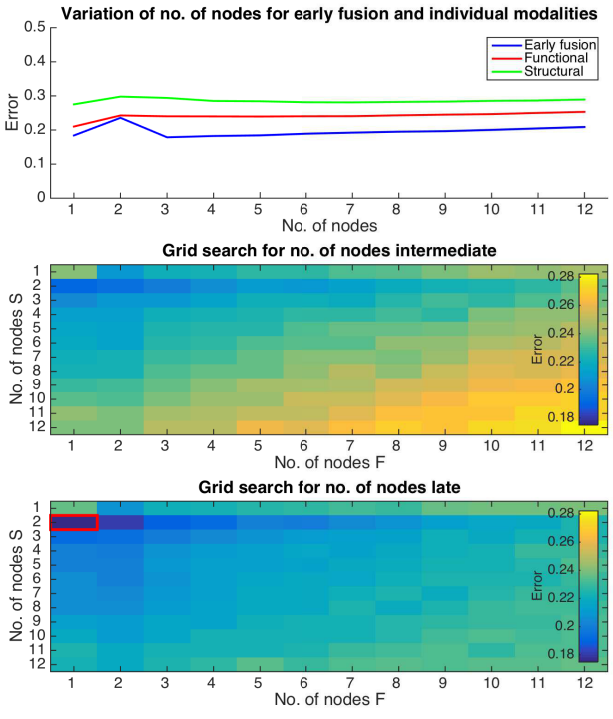


Fig. 3. Grid search for optimal number of nodes for early, intermediate, and late integration of the two modalities, fMRI (F) and sMRI (S), of data set A. A. The red square denotes the best performing node combination in the best performing fusion level (late, F=1, S=2).

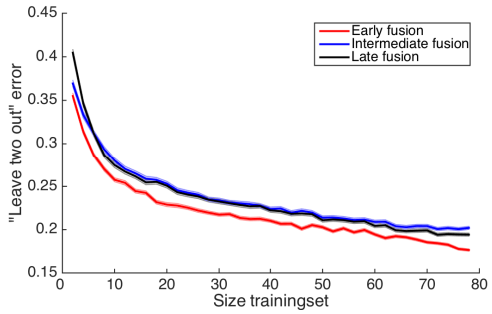


Fig. 4. Learning curve for early, intermediate and late fusion of data set A. The best performing node combinations for each fusion level were analysed for the learning curve. These were Early: 3, Intermediate: (1,2), Late: (1,2).


## Acknowledgment

We gratefully acknowledge the data shared by the Mind Research Network, and funded by a Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the NIH to Dr. Vince Calhoun. The current study was supported by the Lundbeck Foundation (R155-2013-16337).


## References


[1] M. M. Picchioni and R. M. Murray, "Schizophrenia," *BMJ*, vol. 335, no. 7610, pp. 91–95, 2007.

[2] D. D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[3] C. Pohl and J. L. Van Genderen, "Review article Multisensor image fusion in remote sensing: Concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, pp. 823–854, 1998.

[4] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, pp. 345–379, 2010.

[5] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *Journal of Neuroscience Methods*, vol. 204, no. 1, pp. 68–81, 2012.

[6] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2013.

[7] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, pp. 1295–1306, 2009.

[8] M. S. Çetin, F. Christensen, C. C. Abbott, J. M. Stephen, A. R. Mayer, J. M. Cañive, J. R. Bustillo, G. D. Pearlson, and V. D. Calhoun, "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia," *NeuroImage*, vol. 97, pp. 117–126, 2014.

[9] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 97, pp. 273–324, 1997.

[10] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Technical University of Denmark, IMM, Tech. Rep., 2012. [Online]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox

[11] G. E. Hinton, "Training products of experts by minimizing contrastive divergence." *Neural computation*, vol. 14, pp. 1771–1800, 2002.

[12] ——, "A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines," *Computer*, vol. 9, p. 1, 2010.